Mapping gas-phase organic reactivity and concomitant secondary organic aerosol formation: chemometric dimension reduction techniques for the deconvolution of complex atmospheric data sets


1Air Environment Research, School of Environment and Technology, University of Brighton, Brighton, BN2 4GJ, United Kingdom
2Department of Chemistry, University of Leicester, Leicester, LE1 7RH, United Kingdom
3Wolfson Atmospheric Chemistry Laboratories, Department of Chemistry, University of York, York, YO10 5DD, United Kingdom
4School of Earth, Atmospheric and Environmental Sciences, University of Manchester, M13 9PL, United Kingdom
5National Centre for Atmospheric Science, University of Manchester, Manchester, M13 9PL, United Kingdom
6National Centre for Atmospheric Science, University of York, York, YO10 5DD, United Kingdom
7Atmospheric Chemistry Services, Okehampton, Devon, EX20 1FB, United Kingdom
8School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, B15 2TT, United Kingdom
9Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, United Kingdom
10Birmingham Institute of Forest Research, University of Birmingham, Birmingham, B15 2TT, United Kingdom

Correspondence to: K. P. Wyche (k.p.wyche@brighton.ac.uk)

Received: 5 December 2014 – Published in Atmos. Chem. Phys. Discuss.: 20 January 2015
Revised: 14 May 2015 – Accepted: 18 May 2015 – Published: 22 July 2015

Abstract. Highly non-linear dynamical systems, such as those found in atmospheric chemistry, necessitate hierarchical approaches to both experiment and modelling in order to ultimately identify and achieve fundamental process-understanding in the full open system. Atmospheric simulation chambers comprise an intermediate in complexity, between a classical laboratory experiment and the full, ambient system. As such, they can generate large volumes of difficult-to-interpret data. Here we describe and implement a chemometric dimension reduction methodology for the deconvolution and interpretation of complex gas- and particle-phase composition spectra. The methodology comprises principal component analysis (PCA), hierarchical cluster analysis (HCA) and positive least-squares discriminant analysis (PLS-DA). These methods are, for the first time, applied to simultaneous gas- and particle-phase composition data obtained from a comprehensive series of environmental simulation chamber experiments focused on biogenic volatile organic compound (BVOC) photooxidation and associated secondary organic aerosol (SOA) formation. We primarily investigated the biogenic SOA precursors isoprene, α-pinene, limonene, myrcene, linalool and β-caryophyllene. The chemometric analysis is used to classify the oxidation systems and resultant SOA according to the controlling chemistry and the products formed. Results show that “model” biogenic oxidative systems can be successfully separated and classified according to their oxidation products. Furthermore, a holistic view of results obtained across both the gas- and particle-phases shows the different SOA formation chemistry, initiating in the gas-phase, proceeding to govern the differences between the various BVOC SOA compositions. The results obtained are used to describe the particle composition in the context of the oxidised gas-phase matrix. An extension of the technique, which incorporates into the statistical models data from anthropogenic (i.e. toluene) oxidation and “more realistic” plant mesocosm systems, demon-
strates that such an ensemble of chemometric mapping has the potential to be used for the classification of more complex spectra of unknown origin. More specifically, the addition of mesocosm data from fig and birch tree experiments shows that isoprene and monoterpane emitting sources, respectively, can be mapped onto the statistical model structure and their positional vectors can provide insight into their biological sources and controlling oxidative chemistry. The potential to extend the methodology to the analysis of ambient air is discussed using results obtained from a zero-dimensional box model incorporating mechanistic data obtained from the Master Chemical Mechanism (MCMv3.2). Such an extension to analysing ambient air would prove a powerful asset in assisting with the identification of SOA sources and the elucidation of the underlying chemical mechanisms involved.

1 Introduction

Biogenic volatile organic compounds (BVOCs) are ubiquitous in the global troposphere, being emitted primarily from terrestrial plant life (Kanakidou et al., 2005). It is estimated that the total annual emission rate of all (non-methane) BVOCs is roughly 10 times that of all anthropogenic volatile organic compounds (VOCs), being around 750 Tg yr\(^{-1}\) (Sindlarova et al., 2014). With the exception of methane, the most dominant species of BVOCs in terms of emission strength, reactivity and their impact upon the atmosphere are terpenes (Reinnig et al., 2008), a subdivision of BVOCs that primarily comprise the hemiterpene, isoprene (C\(_5\)), monoterpenes (C\(_10\)) and sesquiterpenes (C\(_15\)) (e.g. Atkinson and Arey, 2003a; Kanakidou et al., 2005).

Within the troposphere terpenes are able to react with OH, O\(_3\) and NO\(_3\) at appreciable rates (e.g. Calvert et al., 2000; Koch et al., 2000; Fantechi et al., 2002; Capouet et al., 2004; Kroll et al., 2006) such that their atmospheric lifetimes are of the order of minutes to hours (e.g. Calogirou et al., 1999). Because of their large emission rates and high reactivities, terpenes have a strong impact upon the chemistry of the troposphere at the local, regional and global scales (e.g. Jaoui and Kamens, 2001; Paulot et al., 2012; Surratt, 2013). For instance, terpenes have high photochemical ozone creation potentials (Derwent et al., 2007) and extensive photochemical oxidation pathways that lead to the production of a complex array of oxygenated and nitrated products, some of which are able to form secondary organic aerosol (SOA) (e.g. Calvert et al., 2000; Capouet et al., 2004; Jenkin, 2004; Baltensperger et al., 2008; Kanakidou et al., 2005; Surratt et al., 2006; Kroll and Seinfeld, 2008; Hallquist et al., 2009).

Aerosol particles are natural components of the Earth’s atmosphere responsible for a range of well-documented impacts, ranging from visibility impairment on the local scale to climate change, with suspended particles being able to perturb the Earth’s radiative budget via both direct and indirect mechanisms (Solomon et al., 2007). Furthermore, fine airborne particles have been shown to have numerous detrimental effects on human health, particularly in vulnerable members of the population (Harrison et al., 2010; Heal et al., 2012).

Biogenic SOA (BSOA) has been estimated to account for a significant fraction of total global SOA. Modelling studies suggest the annual global production rate of BSOA is of the order of 16.4 Tg yr\(^{-1}\) (Henze and Seinfeld, 2006). However, despite its importance and the significant amount of investigation conducted upon it, the formation mechanisms and chemical composition of BSOA are still not well characterised (e.g. Librando and Tringali, 2005; Wang et al., 2013). Indeed under certain conditions as much as 80–90 % of analysed SOA mass is unknown (Limbeck et al., 2003; Kalberer et al., 2006). In particular, there remains a significant lack of information regarding the composition and evolution of the complex organic gas-phase matrix during aerosol formation, and its linkage to SOA (Kroll et al., 2005; Librando and Tringali, 2005). Indeed, in the many studies conducted on BSOA, very few gas-phase oxidation products of the precursor are routinely identified and reported.

The chemistry of the atmospheric system is highly nonlinear and can be studied by experiments ranging from highly controlled laboratory studies of a single process, to field studies of the whole complex system. A significant proportion of the findings gained regarding SOA over the last decade and more have come from atmospheric simulation chamber experiments, intermediate in complexity between classical single-process experiments and the fully open system (for various different chamber systems and VOC precursors, see for example, Pandis et al., 1991; Odum et al., 1996; Hoffmann et al., 1997; Griffin et al., 1999; Glasius et al., 2000; Cocker et al., 2001; Jaoui and Kamens, 2003; Kleindienst et al., 2004; Presto et al., 2005; Bloss et al., 2005; Rohrer et al., 2005; Ng et al., 2006, 2007; Dommen et al., 2006; Surratt et al., 2006; Grieshop et al., 2007; Chan et al., 2007; Wyche et al., 2009; Hildebrandt et al., 2009; Rickard et al., 2010; Camredon et al., 2010; Chhabra et al., 2011; Hennigan et al., 2011; Jenkin et al., 2012). Chamber experiments produce a large amount of data, the interpretation of which can often be highly complex and time consuming even though the set-up of the chamber constrains the complexity to a large degree.

In the current “big data” age, advanced monitoring techniques are producing increasingly larger, more complex and detailed data sets. Modern chamber experiments, monitored by state-of-the-art gas- and particle-phase instrumentation, often yield so much data that only a fraction is subsequently used in a given analysis. For example, during a typical 6 h environmental simulation chamber experiment, VOC monitoring chemical-ionisation reaction time-of-flight mass spectrometry (CIR-TOF-MS) will produce roughly 1.1 \(\times\) 10^{7} data points. In order to keep pace with instrument development and maximise the information extracted from
sometimes-complex experiments, it is crucial that we advance our data analysis methods and introduce new data mining techniques.

The work reported here focuses on detailed organic gas-phase and particle-phase composition data, recorded during SOA atmospheric simulation chamber experiments, using CIR-TOF-MS and liquid-chromatography ion-trap mass spectrometry (LC-MS/MS), respectively, as well as broad (i.e. generic composition "type": oxygenated organic aerosol, nitrated, sulfated) aerosol composition data, recorded by compact time-of-flight aerosol mass spectrometry (cTOF-AMS). The goal of this paper is to demonstrate and evaluate the application of an ensemble reductive chemometric methodology for these comprehensive oxidation chamber data sets, to be used as a model framework to map chemical reactivity from mesocosm systems, thus providing a link from model systems to more "real" mixtures of organics. The intermediate complexity offered by simulation chamber experiments makes them an ideal test bed for the methodology. Application of the methodology to resultant particle-phase data also aims to provide a level of particle composition classification in the context of gas-phase oxidation.

Similar approaches using statistical analyses have been recently applied to both detailed and broad ambient aerosol composition data (e.g. Heringa et al., 2012; Paglione et al., 2014), particularly in the context of source apportionment (e.g. Aller et al., 2013). Different methods have been attempted by several groups to deconvolve organic aerosol spectra measured by the aerosol mass spectrometer (AMS) in particular (e.g. Zhang et al., 2005, 2007; Marcollì et al., 2006; Lanz et al., 2007). Zhang et al. (2005) applied a custom principal component analysis (CPCA) method to extract two distinct sources of organic aerosols in an urban environment using linear decomposition of AMS spectra and later applied a multiple-component analysis technique (MCA; an expanded version of the CPCA) to separate more than two factors in data sets from 37 field campaigns in the Northern Hemisphere (Zhang et al., 2007). Marcollì et al. (2006) applied a hierarchical cluster analysis method to an ambient AMS data set, and reported clusters representing biogenic VOC oxidation products, highly oxidised organic aerosols and other small categories. Receptor modelling techniques such as positive matrix factorisation (PMF) employ similar multivariate statistical methods in order to deconvolve a time series of simultaneous measurements into a set of factors and their time-dependent concentrations (Paatero, 1994; Taper, 1994; Paglione et al., 1997). Depending on their specific chemical and temporal characteristics, these factors may then be related to emission sources, chemical composition and atmospheric processing. For example, Lanz et al. (2007) and Ulbrich et al. (2009) applied PMF to the organic fraction of AMS data sets and were able to conduct source apportionment analysis identifying factors contributing to the composition of organic aerosol at urban locations. Slowik et

K. P. Wyche et al.: Mapping gas-phase organic reactivity and concomitant SOA formation
Table 1. Summary of experiments conducted.

<table>
<thead>
<tr>
<th>Experiment ID</th>
<th>Precursor</th>
<th>Structure</th>
<th>k(OH)/k(O₃)³/ cm³ molec⁻¹ s⁻¹</th>
<th>Experiment type (no.)</th>
<th>VOC / NOₓ range</th>
<th>RH/% range</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISOP1 – 10b,c</td>
<td>Isoprene</td>
<td></td>
<td>9.9 × 10⁻¹¹/1.2 × 10⁻¹⁷</td>
<td>Photooxidation (10)</td>
<td>1.3–20.0</td>
<td>49–72</td>
</tr>
<tr>
<td>APIN1 – 4b,c,d,e</td>
<td>α-pinene</td>
<td></td>
<td>5.3 × 10⁻¹¹/8.4 × 10⁻¹⁰</td>
<td>Photooxidation (4)</td>
<td>1.3–2.0f</td>
<td>49–73</td>
</tr>
<tr>
<td>LIM1 – 6b,d,e</td>
<td>Limonene</td>
<td></td>
<td>1.7 × 10⁻¹⁰ / 2.1 × 10⁻¹⁶</td>
<td>Photooxidation (6)</td>
<td>1.4–2.0f</td>
<td>50–82</td>
</tr>
<tr>
<td>BCARY1 – 10b,d,e</td>
<td>β-caryophyllene</td>
<td></td>
<td>2.0 × 10⁻¹⁰/1.2 × 10⁻¹⁴</td>
<td>Photooxidation (10)</td>
<td>0.6–2.0f</td>
<td>50–72</td>
</tr>
<tr>
<td>MYRC1 – 2b,d,e</td>
<td>Myrcene</td>
<td></td>
<td>2.1 × 10⁻¹⁰/4.7 × 10⁻¹⁶</td>
<td>Photooxidation (2)</td>
<td>1.4–1.9</td>
<td>52–54</td>
</tr>
<tr>
<td>LINA1 – 2b,d,e</td>
<td>Linalool</td>
<td></td>
<td>1.6 × 10⁻¹⁰/4.5 × 10⁻¹⁵</td>
<td>Photooxidation (2)</td>
<td>1.4–2.6</td>
<td>42–47</td>
</tr>
<tr>
<td>BIR1 – 2b</td>
<td>Birch trees</td>
<td>Multiple emissions⁸</td>
<td></td>
<td>Mesocosm Photooxidation (2)</td>
<td>5.5–5.6</td>
<td>73–84</td>
</tr>
<tr>
<td>FIG1 – 2b</td>
<td>Fig trees</td>
<td>Multiple emissions⁸</td>
<td></td>
<td>Mesocosm Photooxidation (2)</td>
<td>2.7–9.4</td>
<td>65–75</td>
</tr>
<tr>
<td>TOL1 – 5b</td>
<td>Toluene</td>
<td></td>
<td>3.7 × 10⁻¹²/1/</td>
<td>Photooxidation (5)</td>
<td>1.3–11.6</td>
<td>2–6</td>
</tr>
</tbody>
</table>

ᵃ From Akinos and Arey (2003b); Sun et al. (2012); Khmaganov and Hites (2001) and references therein; b experiments conducted in the MAC; c experiments conducted in the PSIC; d LC-MS/MS filter data available for at least one of these experiments (MAC); e TOF-AMS data available for at least one of these experiments (MAC); f estimated using known volume of reactants injected; g see Wyche et al. (2014); h experiments conducted in the EUHORE.

based on both broad and detailed SOA composition measurements.

The methodology described and the results presented (supported by findings obtained from zero-dimensional box modelling) indicate that there is some potential that the approach could ultimately provide the foundations for a framework onto which it would be possible to map the chemistry and oxidation characteristics of ambient air measurements. This could in turn allow “pattern” typing and source origination for certain complex air matrices and provide a snapshot of the reactive chemistry at work, lending insight into the type of chemistry driving the compositional change of the contemporary atmosphere. There are similarities between this approach to discovery science in the atmosphere and metabolomic strategies in biology (e.g. Sousa et al., 2013; Kuppusami et al., 2015).

2 Experimental details

2.1 Choice of precursors

Six different BVOCs and one anthropogenic VOC were chosen for analysis. The target compounds, their structures and reaction rate constants with respect to OH and O₃ are given in Table 1. The BVOCs were chosen according to their atmospheric prevalence, structure and contrasting photooxidation reaction pathways; all have previously been shown to form SOA under simulation chamber conditions (e.g. Hoffmann et al., 1997; Griffin et al., 1999; Glasius et al., 2000; Jaoui and Kamens, 2003; Presto et al., 2005; Ng et al., 2006; Surratt et al., 2006; Dommen et al., 2006; Lee et al., 2006; Hallquist et al., 2009; Alfarra et al., 2013, and references therein). Isoprene is a C₅ diene that accounts for around 62 % (≈594 Tg yr⁻¹) of total annual non-methane BVOC emissions (Sindelarova et al., 2014). After isoprene, monoterpenes (C₅H₁₀) have the next largest annual emission rate; they account for around 11 % (≈95 Tg yr⁻¹) of total annual non-methane BVOC emissions (Sindelarova et al., 2014). α-pinene and limonene were chosen for analysis here alongside isoprene, the former acting as a model system to represent bicyclic monoterpenes, and the latter to represent monocyclic diene terpenes. In this work, α-pinene and limonene together generically represent (and are referred to hereafter as) “cyclic” monoterpenes (i.e. monoterpenes that contain one six-member carbon ring). In order to explore the chemistry of non-cyclic monoterpenes, myrcene, an acyclic triene monoterpenes, was also included, as was the structurally similar acyclic diene oxygenated VOC (OVOC),
linalool. In this work, myrcene and linalool together generally represent (and are referred to hereafter as) “straight chain” monoterpenes/BVOCs (note: linalool is not technically a monoterpane, but does contain the same carbon backbone as myrcene; consequently, it is expected to exhibit similar photooxidative chemistry). Finally, β-caryophyllene was included to represent sesquiterpenes, which have annual emissions of the order of 20 Tg yr⁻¹ (Sindelarova et al., 2014). In order to test the ability of the methodology to distinguish between biogenic and anthropogenic systems, toluene was also included. Toluene is often used as a model system to act as a proxy for aromatic species in general (Bloss et al., 2005). For contrasting plant mesocosm systems, *Ficus benjamina* and *Ficus cyathistipula* (fig) and *Betula pendula* (birch) species were chosen to represent tropical rainforest and European environs, respectively.

The VOC precursors employed have certain similarities in terms of reaction rate constants with respect to OH and O₃, e.g. limonene, myrcene, linalool and β-caryophyllene all have atmospheric lifetimes with respect to OH of the order of 40–50 min (Alfarra et al., 2013; Atkinson and Arey, 2003b). β-caryophyllene has the shortest lifetime with respect to O₃ (ca. 2 min) and isoprene and α-pinene have the longest lifetimes with respect to both OH and O₃, e.g. isoprene and α-pinene have atmospheric lifetimes with respect to OH of the order of 1.4–2.7 h (Alfarra et al., 2013; Atkinson and Arey, 2003b). In order to ensure the various systems had progressed sufficiently down their respective photooxidative reaction pathways, the experiment duration was set to be sufficiently long so that the majority of the precursor would be consumed by the conclusion of the experiment.

### 2.2 Chamber infrastructure

Experiments were carried out across three different European environmental simulation chamber facilities over a number of separate campaigns. The chambers used, included (1) The University of Manchester Aerosol Chamber (MAC), UK (Alfarra et al., 2012); (2) The European Photoreactor (EUPHORE), ES (Becker, 1996); and (3) The Paul Scherrer Institut Smog Chamber (PSISC), CH (Paulsen et al., 2005). A brief technical description of each facility is given in Table 2.

#### 2.3 Experiment design

Table 1 provides a summary of the experiments conducted, which can be divided into three separate categories: (1) photooxidation, indoor chamber (Wyche et al., 2009; Alfarra et al., 2012, 2013); (2) photooxidation, outdoor chamber (Bloss et al., 2005; Camredon et al., 2010); and (3) mesocosm photooxidation, indoor chamber (Wyche et al., 2014). In each case the reaction chamber matrix comprised a temperature (∼292–299 K) and humidity (49–84 % for photooxidation, indoor chamber and < 2–6 % for photooxidation, outdoor chamber) controlled synthetic air mixture. For all experiments the chamber air matrix also contained a pre-defined initial quantity of NO and NO₂ (VOC / NOₓ ratios in the range 0.6–20, but typical ∼2). The VOC precursor was introduced into the reaction chamber in liquid form via a heated inlet. In the case of the mesocosm photooxidation experiments, a known volume of air containing the precursor VOCs was transferred to the reaction chamber from a separate, illuminated plant chamber, which contained several tree specimens. For the indoor chamber systems, the experiments were initiated, after introduction of all reactants, by the switching on of artificial lights. For the outdoor chamber systems, the opening of the chamber cupola marked the start of the experiment. Experiments were typically run for 4–6 h.

#### 2.4 Instrumentation

CIR-TOF-MS was used to make real-time (i.e. 1 min) measurements of the complex distribution of VOCs (ΣVOC, i.e. the sum of VOCs, oxygenated VOCs – OVOCs and nitrated VOCs – NVOCs) produced in the gas-phase during oxidation of each parent compound. In brief, the CIR-TOF-MS comprises a temperature controlled (T = 40°C) ion source/drift cell assembly coupled to an orthogonal time-of-flight mass spectrometer equipped with a reflectron array (Kore Technology, UK). Proton transfer reaction (PTR) from hydronium (H₃O⁺) and hydrated hydronium (H₃O⁺(H₂O)ₙ) was
employed as the ionisation technique during all experiments (Jenkin et al., 2012; Wyche et al., 2007). Further details regarding the CIR-TOF-MS can be found in Blake et al. (2004) and Wyche et al. (2007).

Aerosol samples were collected on 47 mm quartz fibre filters at the end of certain experiments and the water-soluble organic content was extracted for analysis using LC-MS/MS. Reversed-phase LC separation was achieved using an HP 1100 LC system equipped with an Eclipse ODS-C18 column with 5 µm particle size (Agilent, 4.6 mm x 150 mm). Mass spectrometric analysis was performed in negative ionisation mode using an HCT-Plus ion trap mass spectrometer with electrospray ionisation (Bruker Daltonics GmbH). Further details can be found in Hamilton et al. (2011).

For several experiments, real-time broad chemical characterisation of the SOA was made using a cTOF-AMS (Aerodyne Research Inc., USA). The cTOF-AMS was operated in standard configuration, taking both mass spectrum (MS) and particle time-of-flight (PTOF) data; it was calibrated for ionisation efficiency using 350 nm monodisperse ammonium nitrate particles, the vapouriser was set to ∼ 600 °C and a collection efficiency value of unity was applied (Alfarra et al., 2006). For further details, refer to Drewnick et al. (2005) and Canagaratna et al. (2007).

Filter and cTOF-AMS data were collected only during photooxidation experiments conducted at the MAC. Repeat experiments conducted at the MAC were carried out under similar starting conditions (e.g. VOC/NOx ratio; Alfarra et al., 2013).

Each chamber was additionally instrumented with online chemiluminescence/photolytic NOx analysers, UV photometric O3 detectors, and scanning mobility particle sizers and condensation particle counters for aerosol size and number concentration, as well as temperature, pressure and humidity monitors. For full details regarding the various instrument suites employed at each chamber see Alfarra et al. (2012), Paulsen et al. (2005), Camredon et al. (2010) and references therein.

2.5 Model construction

In order to aid analysis, the composition and evolution of the gas-phase components of the α-pinene chamber system were simulated using a chamber optimised photochemical box model incorporating the comprehensive α-pinene atmospheric oxidation scheme extracted from the Master Chemical Mechanism website (Jenkin et al., 1997, 2012; Saunders et al., 2003; http://mcm.leeds.ac.uk/MCM). The α-pinene mechanism employed (along with an appropriate inorganic reaction scheme) contained approximately 313 species and 942 different reactions. The box model employed also incorporated a series of “chamber-specific” auxiliary reactions adapted from Bloss et al. (2005), Zador et al. (2006) and Metzger et al. (2008) in order to take into account background chamber reactivity. Photolysis rates were parameterised for the PSI chamber and constrained using measured values of (j(NO2)). All simulations were run at 295 K and 50 % relative humidity. NO, NO2, HONO and α-pinene were either initialised or constrained, depending on the scenario investigated. For further details regarding the chamber auxiliary mechanism and similar modelling exercises, see for example Rickard et al. (2010).

3 Data analysis

3.1 Data processing

All CIR-TOF-MS data were recorded at a time resolution of 1 min. In order to remove the time dimension and simultaneously increase detection limit, the individual mass spectra were integrated over the entire experiment; as such, no account is taken of overall reaction time in the CIR-TOF-MS analysis. Removing the time dimension acts to reduce the dimensionality of the data, whilst maintaining the central characteristic spectral fingerprints produced by the photooxidation process. On average across all experiments studied, 98 % of the precursor had been consumed by the conclusion of the experiment; hence, it is assumed that sufficient reaction took place in each instance to provide summed-normalised mass spectra that fully capture first- and higher-generation product formation.

The resultant summed spectra were normalised to 106 primary reagent ion counts (i.e. Σ(H2O+ + H3O+ + (H2O)n)). Similarly normalised background spectra (recorded prior to injection of the precursor) were then subtracted from the summed-and-normalised experiment spectra. The 65 < m/z < 255 channels of the background removed spectra were extracted to comprise the region of interest. These ions tend to carry the most analyte-specific information, with lower m/z features tending to comprise either generic fragment ions that provide little chemical information (Blake et al., 2006) and/or small “common” compounds, some of which can be emitted from illuminated chamber walls (e.g. Bloss et al., 2005; Zador et al., 2006; Metzger et al., 2008).

The extracted data were refined further by the application of a Mann–Whitney test (see Statistical Analysis for details), leaving residual spectra that comprised only the integrated-over-time signals corresponding to the VOC precursor and any product VOCs formed within the chamber during the experiment. Finally, the signal counts (in units of normalised counts per second; ncps) in each mass channel of the residuals, were expressed as a percentage of the total ion count in the refined region of interest.

The LC-MS/MS signal intensity data for the region 51 < m/z < 599 were extracted for analysis. For the AMS data, a 10 min average was produced at 4 h after lights on (hour-4; roughly around the time when SOA mass had reached a peak and towards to the end of the experiment) and the region 40 < m/z < 150 (again the region carrying the
most information; Alfarra et al., 2006) was extracted. Similar to the gas-phase data sets, the LC-MS/MS and CTOF-AMS data were filtered using a Mann–Whitney test. Finally, for each data set all signal counts were expressed as a percentage of the total ion count in the respective \( m/z \) region of interest.

### 3.2 Statistical analysis

Before any multivariate analysis was conducted, the processed CIR-TOF-MS, LC-MS/MS and CTOF-AMS spectra were first filtered to remove unwanted data that were deemed not to be statistically significant. In order to do this, the mass spectra were initially grouped by structure of the precursor employed, giving seven separate groups for the CIR-TOF-MS data and three groups (owing to the smaller number of precursor species investigated) for the LC-MS/MS and CTOF-AMS data, respectively. A two-sided Mann–Whitney test was then used to assess whether signals reported in individual mass channels were significantly different from the corresponding signals measured during a blank experiment. SPSS V20 (IBM, USA) was used for the analysis. A \( p \) value of < 0.05 was considered statistically significant. The final summed-normalised and filtered spectra were then subjected to a series of multivariate statistical analysis techniques in order to probe the underlying chemical information. PLS-Toolbox (Eigenvector Research Inc., USA), operated in Matlab (Mathworks, USA; PLS-Tool Box), was used for the analysis.

To begin with, to reduce the data and identify similarities between the precursor oxidation systems, a PCA was conducted on the BVOC data set and the model generated was then employed to map the reactivity of fig and birch tree mesocosm systems and to investigate the fit of a typical anthropogenic system (toluene) into the PCA space (both introduced into the model as test data sets). An unsupervised pattern recognition, hierarchical cluster analysis was also conducted on the data and a dendrogram was produced to test relatedness, support the PCA and help interpret the precursor class separations achieved. The dendrogram was constructed using PCA scores, the centroid method and Mahalanobis distance coefficients. Finally, a supervised pattern recognition PLS-DA analysis was employed as a check for false positives and as a quantitative classification tool to test the effectiveness of classification of the various systems in the model.

For the superposition of “classification” confidence levels onto the results of the PCA and HCA and for classification discrimination in the PLS-DA, prior to analysis the experiments were grouped according to the structure of the precursor investigated. Group 1 = isoprene (hemiterpene) and group 2 = \( \alpha \)-pinene and limonene (both cyclic monoterpenes with an endocyclic double bond). Although limonene also has an exocyclic double bond in a side chain, we justify this classification on account of the endocyclic double bond in limonene being much more reactive towards ozone and slightly more reactive towards OH (Calvert et al., 2000). Group 3 = \( \beta \)-caryophyllene (sesquiterpene) and group 4 = myrcene (straight chain monoterpane) and linalool (straight chain OVOC). Strictly speaking, linalool is an OVOC (structure \( \text{C}_{10}\text{H}_{18}\text{O} \)) and not a monoterpane (structure \( \text{C}_{10}\text{H}_{16} \)); however we justify this grouping on account of both myrcene and linalool comprising primary BVOCs (often co-emitted; Bouvier-Brown et al., 2009; Kim et al., 2010; Wyche et al., 2014) with certain structural similarities.

### 4 Results

#### 4.1 Experiment overview

The temporal evolution of various key gas-phase (a) and particle-phase (b) parameters measured during a typical photooxidation experiment, are shown in Fig. 1 in order to provide background context. In this instance the precursor was myrcene and the facility employed was the MAC. Full details describing the underlying chemical and physical mechanisms at play within such experiments can be found elsewhere (e.g. Larsen et al., 2001; Bloss et al., 2005; Paulsen et al., 2005; Surratt et al., 2006, 2010; Wyche et al., 2009, 2014; Camredon et al., 2010; Rickard et al., 2010; Eddingsaas et al., 2012a and b; Hamilton et al., 2011; Jenkin et al., 2012; Alfarra et al., 2012, 2013, and references therein).

#### 4.2 Mapping gas-phase composition

Of the 191 different mass channels extracted from the CIR-TOF-MS data for analysis (i.e. \( 65 < m/z < 255 \)), the Mann–Whitney test identified 151 as significant for one or more of the terpene precursor groups tested. These data were subsequently subjected to PCA. From inspection of the Eigenvalues derived, four principal components (PCs) were selected for analysis, which collectively accounted for 96% of the variance within the data, with PCs 1 and 2 accounting for the vast majority, i.e. 63 and 18%, respectively. This step, therefore, reduced the temporal traces of 191 mass spectrum peaks to four composite and orthogonal dimensions.
Figure 2 shows a loadings bi-plot of PC2 vs. PC1. It is clear from Fig. 2, that the model is able to successfully separate the four different classes of biogenic systems investigated. β-caryophyllene mass spectra are grouped in the upper left-hand quadrant of Fig. 2, the monoterpenes in the lower left-hand quadrant and isoprene to the centre right. Moreover, the principal component analysis is able to distinguish between the cyclic monoterpene experiments of limonene and α-pinene (grouped into one class), and the straight chain monoterpene experiments of myrcene and linalool (grouped into a second class), albeit with the latter having a greater spread in confidence.

The m/z loadings of the PCA allow us to understand how the spectral fingerprints of the different terpene oxidation systems are grouped/separated by the PCA model. The first set of ions that contribute to separation of the different terpene systems comprises the protonated parent ions (MH⁺) of the precursors themselves (and major fragments thereof), i.e. m/z 69 for isoprene, 137 (and fragment 81) for all monoterpenes (regardless of structure) and 205 for β-caryophyllene. Important contributions are to be expected from the respective parent ions (being the basis for the use of chemical-ionisation mass spectrometry as an analyser of gas mixtures; Blake et al., 2009). Our purpose here goes beyond identification of precursor and intermediate VOCs to an interpretation of reaction pathways in complex mixtures and potential linkages to SOA. In doing this, a certain amount of disambiguation of isobaric compounds becomes possible; indeed, as discussed in more detail below, Fig. 2 clearly shows separation between cyclic and non-cyclic monoterpene oxidation groups, both of which have precursors of molecular weight (MW) 136 g mol⁻¹. Note, for clarity within Fig. 2, the scale has been set to show the bulk of the data; hence, precursor parent ions and m/z 71 are not shown.

Moving past the precursors into the detailed chemical information provided by the oxidation products formed within the chamber, we can see from the data and Fig. 2 that amongst others, m/z 71 (methyl vinyl ketone and methacrolein), 72 (methyl vinyl ketone and methacrolein C13 isotope), 75 (hydroxy acetone), 83 (C₅-hydroxy carbonyls/3-methyl furan) and 87 (C₄-hydroxy carbonyls/methacrylic acid) all contribute to separation of the isoprene group, and m/z 237 (β-caryophyllene aldehyde) and 235 and 253 (β-caryophyllene secondary ozonide and isomers thereof) to that of the β-caryophyllene group. The monoterpene groupings are influenced by the
presence of m/z 107, 151 and 169 (primary aldehydes, pinonaldehyde and limononaldehyde) and 139 (primary ketone, limonaketone) ions in their mass spectra. Helping to separate the straight chain from cyclic monoterpenes are m/z 95 and 93, relatively dominant features in both the myrcene and linalool spectra (relative abundance 10–24 % for m/z 93). m/z 93 has previously been identified as a major fragment ion of first generation myrcene and linalool products 4-vinyl-4-pentenal and 4-hydroxy-4-methyl-5-hexen-1-ol, respectively (Shu et al., 1997; Lee et al., 2006). A list of major ions contributing to the separation of spectra in statistical space is given in Table 3 along with potential identities and precursors. It is worthy of note here that these ions and the overall fragmentation patterns observed in this study are largely in-line with those reported by Lee et al. (2006), in their comprehensive PTR-MS analysis of a wide range of BVOC precursors and their associated oxidation products.

4.3 Implementation of the model to classify mesocosm data

Having employed the terpene data as a training set to construct a PCA model, a test set of mesocosm data was introduced in order to investigate the ability of the model to map the classification of more complex biogenic mixtures. In this instance the mesocosm test set comprised two birch tree and two fig tree photooxidation experiments, containing a more complex and “realistic” mixture of various different VOCs (Wyche et al., 2014). The resultant scores plot is shown in Fig. 3.

Figure 3 demonstrates that the model can successfully distinguish between the two different types of mesocosm systems. Moreover, the model correctly classifies the mesocosm systems within the PCA space, with the birch trees, which primarily emit monoterpenes and only small quantities of isoprene (Wyche et al., 2014), grouped with the single-precursor monoterpenes cluster and the fig trees, which primarily emit isoprene and camphor and only a small amount of monoterpenes (Wyche et al., 2014), grouped between the monoterpene and isoprene clusters. Investigation of the mesocosm mass spectra and PCA loadings shows that mass channels 137, 139, 107, 95, 93, 81 and 71 are amongst features important in classifying the birch tree systems, with the relatively strong presence of m/z 93 suggesting the emission of non-cyclic as well as cyclic monoterpenes from the birch trees. This was confirmed by cross reference with GC-MS analysis, which showed that the acyclic monoterpene, ocimene, was the third most abundant monoterpenes present in the birch tree emissions (Wyche et al., 2014). For the fig tree systems, mass channels 153, 81, 73, 71 and 69 are key for classification, with the presence of small quantities of camphor (m/z 153) and monoterpenes (m/z 81) causing the group to undergo a lateral shift in the PCA space, along PC1 away from the single-precursor isoprene cluster.

As a further test of the technique to distinguish between and to classify VOCs and the matrix of oxidised organic compounds that may derive from their atmospheric chemistry, test data from an anthropogenic system was introduced into the model. In this instance, the toluene photooxidation system was employed. Toluene is an important pollutant in urban environments, originating for example from vehicle exhausts and fuel evaporation; furthermore, it represents a model mono-aromatic, SOA precursor system (e.g. Bloss et al., 2005). As can be seen from the resultant scores plot in Fig. 4, the model is also able to discriminate the anthro-
Table 3. List of certain major product ions integral to the separation of BVOC photooxidation spectra in statistical space, their corresponding tentative assignments and their precursor. See main text (Sects. 4.2 and 4.5) for further information.

<table>
<thead>
<tr>
<th>Ion / m/z</th>
<th>CIR-TOF-MS</th>
<th>LC-MS/MS</th>
<th>cTOF-AMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignment</td>
<td>Precursor</td>
<td>Assignment</td>
<td>Precursor</td>
</tr>
<tr>
<td>41 –</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>43 –</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>44 –</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>71 methyl vinyl ketone + methacrolein</td>
<td>isoprene</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>75 hydroxy acetone</td>
<td>isoprene</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>83 ( \text{C}_3\text{-hydroxy carbonyls} / 3\text{-methyl furan} )</td>
<td>isoprene</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>87 ( \text{C}_4\text{-hydroxycarbonyls} / 3\text{-methyl furan} )</td>
<td>isoprene</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>93 4-vinyl-4-pentenal / 4-hydroxy-4-methyl-5-hexen-1-al</td>
<td>myrcene / linalool</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>107 piononaldehyde / limononaldehyde</td>
<td>( \alpha )-pinene / limonene</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>111 4-vinyl-4-pentenal / 4-hydroxy-4-methyl-5-hexen-1-al</td>
<td>myrcene / linalool</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>139 limonaketone</td>
<td>limonene</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>151 piononaldehyde / limononaldehyde</td>
<td>( \alpha )-pinene / limonene</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>169 piononaldehyde / limononaldehyde</td>
<td>( \alpha )-pinene / limonene</td>
<td>pinalic 3-acid / ketolimonon-alddehyde + limonalic acid</td>
<td>( \alpha )-pinene / limonene</td>
</tr>
<tr>
<td>183 –</td>
<td>–</td>
<td>pinonic acid / limononic acid + 7-hydroxylimonon-alddehyde</td>
<td>( \alpha )-pinene / limonene</td>
</tr>
<tr>
<td>185 –</td>
<td>–</td>
<td>pinic acid / limonic acid</td>
<td>( \alpha )-pinene / limonene</td>
</tr>
<tr>
<td>199 –</td>
<td>–</td>
<td>( \text{C}<em>9\text{H}</em>{13}\text{O}_5 )</td>
<td>myrcene</td>
</tr>
<tr>
<td>215 –</td>
<td>–</td>
<td>( \text{C}<em>{10}\text{H}</em>{15}\text{O}_3 )</td>
<td>myrcene</td>
</tr>
<tr>
<td>227 –</td>
<td>–</td>
<td>( \text{C}<em>{10}\text{H}</em>{17}\text{O}_6 )</td>
<td>myrcene</td>
</tr>
<tr>
<td>235 ( \beta )-caryophyllene secondary ozonide + isomers</td>
<td>( \beta )-caryophyllene</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>237 ( \beta )-caryophyllene aldehyde</td>
<td>( \beta )-caryophyllene</td>
<td>3-[2,2-dimethyl-4-(1-methylene-4-oxo-buty1)-cyclobutyl]propanoic acid</td>
<td>( \beta )-caryophyllene</td>
</tr>
<tr>
<td>251 –</td>
<td>–</td>
<td>( \beta )-caryophyllonic acid</td>
<td>( \beta )-caryophyllene</td>
</tr>
<tr>
<td>253 ( \beta )-caryophyllene secondary ozonide + isomers</td>
<td>( \beta )-caryophyllene</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>255 –</td>
<td>–</td>
<td>4-(2-(2-carboxyethyl)-3,3-dimethylcyclobutyl)-4-oxobutanoic acid</td>
<td>( \beta )-caryophyllene</td>
</tr>
<tr>
<td>267 –</td>
<td>–</td>
<td>( \beta )-14-hydroxy-caryophyllonic acid +( \beta )-10-hydroxy-caryophyllonic acid</td>
<td>( \beta )-caryophyllene</td>
</tr>
<tr>
<td>271 –</td>
<td>–</td>
<td>4-(2-(3-hydroperoxy-3-oxopropyl)-3,3-dimethylcyclobutyl)-4-oxobutanoic acid / 4-(2-(2-carboxy-1-hydroxyethyl)-3,3-dimethylcyclobutyl)-4-oxobutanoic acid</td>
<td>( \beta )-caryophyllene</td>
</tr>
<tr>
<td>321 –</td>
<td>–</td>
<td>( [\text{M-H}_2^{+}\text{FA}^{+}\text{Na}^-] )</td>
<td>myrcene</td>
</tr>
</tbody>
</table>
Figure 5. Dendrogram showing the grouping relationship between the various gas-phase matrices of systems examined. Red – isoprene, pink – fig, green – cyclic monoterpenes (α-pinene and limonene), yellow – birch, light-blue – straight chain monoterpenes (myrcene and linalool) and dark-blue – sesquiterpene (β-caryophyllene).

The relationships between the various terpene and mesocosm systems and their groupings with respect to one another can be explored further via the implementation of HCA; Fig. 5 gives the dendrogram produced. Inspection of Fig. 5 provides further evidence that the various systems in the four classes of terpenes investigated distinctly group together, with overall relatedness < 1 on the (centroid) distance between clusters scale using the Mahalanobis distance measure (Mahalanobis, 1936). Figure 5 shows that the sesquiterpene oxidation system has the most distinct spectral fingerprint (containing distinctive, higher mass oxidation products, e.g. m/z 253) and that the cyclic and straight chain monoterpenes systems appear the most similar (with some common features alongside key, unique precursor/mechanism specific product patterns, e.g. m/z 93 for myrcene and linalool), grouping together with subclusters of cyclic and non-cyclic precursors. The monoterpenes dominated birch tree mesocosm experiments are grouped with the cyclic monoterpenes and show a close relationship with non-cyclic monoterpenes systems. Being dominated by isoprene emissions, yet with some monoterpenes and camphor present, the fig tree mesocosm experiments group separately but with a relatively close degree of relation to the single-precursor isoprene experiments.

In order to advance our chemometric mapping of biogenic systems beyond PCA and HCA (which do not consider user supplied a priori observation “class” information) and to provide a degree of quantification to our analysis, a PLS-DA using six latent variables (LVs) was conducted on the terpene and mesocosm data. For the PLS-DA, the experiments were grouped into their respective “classes”, i.e. hemiterpene – isoprene; cyclic monoterpene – α-pinene and limonene; sesquiterpene – β-caryophyllene; non-cyclic monoterpene – myrcene and linalool; birch trees; fig trees. Figure 6 shows a plot of the resultant scores on the first three LVs (accounting for ~85 % of the variance), from which it is clear that the PLS-DA is able to successfully discriminate between the four terpene classes, and places the monoterpenes dominant birch
experiments within the single-precursor monoterpene cluster, and the isoprene dominant fig experiments close to the single-precursor isoprene cluster within the statistical space. The greater spread in confidence of the non-cyclic monoterpene group is likely to result, to some extent, from the low number of repeat experiments employed (i.e. only two each for myrcene and linalool).

As can be seen from inspection of Table 4, model classification sensitivity and specificity was high in each instance. Each of the biogenic systems studied were predicted with 100% sensitivity (with the exception of birch mesocosm), meaning that each set of experiments (again, except birch mesocosm) was predicted to fit perfectly within its class. The relatively low sensitivity obtained for birch mesocosm (50%), is most likely a result of the use of only two repeat experiments in the model, coupled with experiment limitations and ageing trees producing slightly lower emissions during the final birch mesocosm experiment. All of the systems were predicted with >90% specificity (four of the six with 100% specificity), indicating that all experiments are highly unlikely to be incorrectly classified.

4.5 Mapping particle-phase composition

In order to explore similar classifications and linkages in the concomitant particle-phase, the PCA, HCA and PLS-DA techniques were also applied to the off-line LC-MS/MS spectra obtained from analysis of filter samples and on-line cTOF-AMS spectra.

As can be seen from inspection of Fig. 7, the detailed LC-MS/MS aerosol spectra produce PCA results somewhat similar to those of the gas-phase CIR-TOF-MS spectra, with distinct clusters of cyclic monoterpenes, straight chain monoterpenes and sesquiterpenes. From inspection of the loadings components of the bi-plot (Fig. 7a), we can see that m/z 237 (3-[2,2-dimethyl-4-(1-methylene-4-oxobutyl)-cyclobutyl]-propanoic acid), 251 (β-caryophyllonic acid), 255 (4-(2-(2-carboxyethyl)-3,3-dimethylcyclobutyl)-4-oxobutanoic acid), 267 (β-14-hydroxycaryophyllonic acid and β-10-hydroxycaryophyllonic acid) and 271 (4-(2-(3-hydroperoxy-3-oxopropyl)-3,3-dimethylcyclobutyl)4-oxobutanoic acid or 4-(2-(2-carboxy-1-hydroxyethyl)-3,3-dimethylcyclobutyl)-4-oxobutanoic acid) are amongst those ions dominant in classifying the sesquiterpenes. For further details regarding β-caryophyllene oxidation products, see for example Lee et al. (2006), Winterhalter et al. (2009), Hamilton et al. (2011), Chan et al. (2011), Li et al. (2011) and Jenkin et al. (2012) and references therein, as well as Sect. 5. Of this set of oxidation products, β-caryophyllonic acid is common between the gas (i.e. m/z 253) and particle (i.e. m/z 251) phases.

Similarly, those ions (compounds) significant in isolating the cyclic monoterpenes include, m/z 169 (pinalic-3-acid, ketolimononaldehyde and limonic acid), 183 (pinonic acid, limononic acid and 7-hydroxylimononaldehyde) and 185 (pinic acid and limonic acid), of which only those compounds of m/z 169 were observed to be of significant contribution to the gas-phase composition (observed as m/z 171; relative contribution as high as 1–5% during α-pinene experiments). For further details regarding α-pinene and limonene oxidation products, see for example Larsen et al. (2001), Jaoui et al. (2003, 2006), Capouet et al. (2004), Jenkin (2004), Lee et al. (2006), Ng et al. (2006), Camredon et al. (2010) and Hamilton et al. (2011) and references therein as well as Sect. 5. Comparatively little information is available on the speciated composition of myrcene and linalool SOA; however from Fig. 7a it is clear that here, compared to the cyclic monoterpenes, somewhat larger mass compounds are important in classifying straight chain monoterpenes, e.g. m/z 321 (adduct ion [M-H2+ FA + Na]+ M = 254 Da; potential formulae – C12H14O6, six double bond equivalents or C13H18O5, five double bond equivalents; indicative of oligomer formation), 325, 322 (the C13 peak for the m/z 321 ion), 227 (C10H11O6), 215 (C10H13O5) and 199 (C9H11O5). Compounds of such high molecular weight were not observed in the concomitant gas-phase spectra. A list of major ions contributing to the separation of spectra in statistical space is given in Table 3 along with potential identities and precursors.

As with the PCA, the dendrogram produced via cluster analysis of the LC-MS/MS particle-phase data gave three distinct clusters (Fig. 7b), i.e. cyclic monoterpene, straight chain monoterpenes and sesquiterpene. The corresponding PLS-DA analysis reported 100% sensitivity in each case and 100% specificity for all systems except sesquiterpenes (i.e. β-caryophyllene = 83%), suggesting a good level of model classification for the three types of terpene systems studied.

Figure 6. Scores plot of the first three latent variables derived from the PLS-DA model analysis of the isoprene, cyclic monoterpene (α-pinene and limonene), sesquiterpene (β-caryophyllene), straight chain monoterpenes (myrcene and linalool), fig and birch chamber data. Classification confidence level is 95%.
Table 4. PLS-DA model classification sensitivity and specificity for the gas-phase biogenic air matrices.

<table>
<thead>
<tr>
<th>Cross-validation</th>
<th>Isoprene</th>
<th>Cyclic-monoterpene</th>
<th>Sesquiterpene</th>
<th>Straight-chain-monoterpene</th>
<th>Fig tree</th>
<th>Birch tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity (%)</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>100.0</td>
<td>92.9</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>91.7</td>
</tr>
</tbody>
</table>

Figure 7. (a) Loadings bi-plot of the second vs. first principal components obtained from the PCA of LC-MS/MS aerosol spectra from a subset of terpene experiments and (b) the corresponding HCA dendrogram. See main text (Sect. 4.5 and Table 3) for further information, including ion assignments.

Despite utilising the somewhat destructive electron impact (EI) ionisation technique, the cTOF-AMS produces spectra of sufficient chemical detail such that the PCA and HCA are able to successfully differentiate between the groups of terpenes tested (Fig. 8a and b). However, unlike the outputs from the CIR-TOF-MS and LC-MS/MS PCAs, the cyclic and straight chain monoterpenes in the cTOF-AMS PCA do not group into two distinct classes, instead they tend to group in their species-specific sub-classes within the upper region of the PCA space. Indeed, the PLS-DA gave 100% sensitivity and specificity for the cyclic monoterpenes and sesquiterpenes, but only 75% sensitivity for the straight chain monoterpenes, suggesting that the model does less well at assigning myrcene and linalool cTOF-AMS spectra to their defined class.

As can be seen from inspection of Fig. 8a, α-pinene, limonene and linalool tend in general to cluster towards the upper and right regions of the PCA space, primarily owing to the significant presence of m/z 43 in their spectra; both ions constituting common fragments observed in AMS of SOA (Alfarra et al., 2006). During such chamber experiments, the m/z 43 peak tends to comprise the OOCO⁺ ion, originating from oxidised compounds containing carbonyl functionalities; it is usually representative of freshly oxidised material and semi-volatile oxygenated organic aerosol (SV-OOA; Alfarra et al., 2006, 2013).

From further inspection of the loadings bi-plot (Fig. 8a), we see that the four sesquiterpene (β-caryophyllene) experiments cluster towards the lower left-hand quadrant, their clustering heavily influenced by the presence of m/z 41 in their spectra as well as m/z 55, 79 and 95. In EI-AMS, m/z 41 comprises the unsaturated C₅H₅⁺ fragment (Alfarra et al., 2006). As well as being influenced by the m/z 41 ion, the myrcene cluster (situated in the region of both the α-pinene and β-caryophyllene clusters in the PCA space) is also influenced by m/z 44, i.e. most likely the CO₂⁺ ion. In this instance m/z 44 would tend to result from low volatility oxygenated organic aerosol (LV-OOA), derived from highly oxidised compounds, including oxo- and di-carboxylic acids (Alfarra et al., 2004, 2006). For full details regarding the particle-phase-specific experiments conducted at the MAC, see Hamilton et al. (2011), Jenkin et al. (2012) and Alfarra et al. (2012, 2013).

5 Discussion

5.1 Mapping chemistry

Figure 9 provides a highly simplified overview of the current state of knowledge regarding the atmospheric oxidation
of hemi-, sesqui-, cyclic and straight chain monoterpenes, showing selected key steps and intermediates on route to SOA formation. The mechanisms outlined in Fig. 9 underpin the findings reported here and help to explain how the atmospheric chemistry of the various terpene oxidation systems and their SOA can be chemometrically mapped with respect to one another.

From a review of recent literature and from the summary presented in Fig. 9, it can be seen that isoprene can react to form condensable second and higher-generation nitrates in the presence of NO\textsubscript{x}, e.g. C\textsubscript{4}-hydroxy nitrate peroxy acetyl nitrate (C4-HN-PAN in Fig. 9) (Surratt et al., 2010), as well as condensable OVOCs, e.g. hydroxymethyl-methyl-\alpha-lactone (HMML) (Kjaergaard et al., 2012) and methacrylic acid epoxide (MAE) (Lin et al., 2013), via methacrolein (MACR) and methacryloyl-peroxy nitrate (MPAN). Alternatively, under “low NO\textsubscript{x}” conditions (e.g. < 1 ppbV) isoprene can react to form condensable second-generation epoxides, e.g. isoprene epoxides (IEPOX), via primary peroxides (ISOPOOH) (Paulot et al., 2009a; Surratt et al., 2006). Such C\textsubscript{4} and C\textsubscript{5} saturated, low volatility species constitute the monomer building blocks that proceed to form relatively high O:C ratio (nitrated in the presence of NO\textsubscript{x} and sulfated in the presence of H\textsubscript{2}SO\textsubscript{4}) isoprene SOA oligomers (e.g. 2-methyl tetrodimer, O:C = 7:9) (Claeys et al., 2004; Surratt et al., 2006, 2010; Worton et al., 2013). Consequently, the gas-phase composition under conditions forming isoprene SOA will therefore be dominated by relatively low MW monomer precursors, e.g. MACR (MH\textsuperscript{+} = m/z 71 in the CIR-TOF-MS spectra), isoprene nitrates (ISOPOH in Fig. 9; MH\textsuperscript{+}-HNO\textsubscript{3} = m/z 85) and MPAN (MH\textsuperscript{+}H\textsubscript{2}O-HNO\textsubscript{3} = m/z 103) under “high NO\textsubscript{x}” conditions (e.g. ∼10s–100s ppbV (Paulot et al., 2009b; Surratt et al., 2006, 2010), and ISOPOOH and IEPOX (MH\textsuperscript{+}, H\textsubscript{2}O = m/z 101) under “low NO\textsubscript{x}” conditions. For the “high NO\textsubscript{x}” isoprene experiments conducted here, besides m/z 71, i.e. MACR (measured together with methyl vinyl ketone), m/z 87, 85, 83 and 75, i.e. (tentatively assigned to be) C\textsubscript{4}-hydroxy carbonyls/methacrylic acid, ISOPOH, C\textsubscript{5}-hydroxy carbonyls (C5HC in Fig. 9)/3-methyl furan (3-MF) and hydroxy acetone, respectively, were significant in classifying the isoprene group; MPAN at the m/z 103 ion was only a minor contributor. It should be noted that in theory, both HMML and MAE (MH\textsuperscript{+} = m/z 103) may produce fragment ions of m/z 85 (i.e. MH\textsuperscript{+}-H\textsubscript{2}O) following PTR ionisation; however without further detailed characterisation we are unable at this stage to postulate their fractional contribution to the measured m/z 85 signal.

Depending on the chemistry involved (Fig. 9), potential SOA forming monoterpane products will either be (six-member) ring-retaining (e.g. from reaction with OH) or (six-member) ring cleaved (e.g. from reaction with OH or O\textsubscript{3}), leading to gas-phase spectra with mid-MW C\textsubscript{9} and C\textsubscript{10} oxygenated (and nitrated in the presence of NO\textsubscript{x}) products (e.g. Kamens and Jaoui, 2001; Larsen et al., 2001; Capouet et al., 2004; Yu et al., 2008; Camredon et al., 2010; Eddingsaas et al., 2012a and b). Both (six-member) ring-retaining and (six-member) ring-opening products have been observed in monoterpane SOA (e.g. Yu et al., 1999; Larsen et al., 2001; Camredon et al., 2010), with the latter generally being dominant in terms of abundance (Camredon et al., 2010). Furthermore, (six-member) ring-opening products are believed to undergo chemistry within the aerosol to form relatively low O:C ratio oligomers (e.g. 10-hydroxy-pinonic acid-pinonic acid dimer, O:C = 7:19) (Gao et al., 2004; Tolocka et al., 2004; Camredon et al., 2010).

OH will react with straight chain monoterpenes, such as myrcene, primarily by addition to either the isolated or the conjugated double bond system. Reaction at the isolated C=C bond can proceed via fragmentation of the carbon backbone, producing acetone and mid-MW, unsaturated C\textsubscript{7} OVOCs and/or NVOCs, depending on NO\textsubscript{x} levels). React-
Figure 9. Simplified schematic illustrating some of the important mechanistic pathways in the gas-phase oxidation of isoprene, α-pinene, β-caryophyllene and myrcene, and the associated mass transfer to the particle phase (obtained from the current literature; see text, Sect. 5 for references). Red arrows and text represent “high NOx” pathways, green arrows and text – ozonolysis reactions, grey arrow and text – speculative and dashed arrows – multiple steps. * represents multiple photooxidative routes initiated by reaction with OH (i.e. involving the reactants – HO, O, NO, HO2 and/or RO2), leading to structurally similar products containing different functional groups. α-pinene mechanism – X = OH, =O, OOH or ONOO2; Y = CHO or C(O)OH; Z = OH, OOH or ONOO2; β-caryophyllene mechanism – X = CH2OH(OH), CH2OH(OOH), CH2OH(ONO2) or = O. Myrcene mechanism – Y = OOH or ONOO2; Z = CHO or C(O)OH.
tion at the conjugated double bond system in myrcene would be expected to form formaldehyde in conjunction with either a C₉ aldehyde or C₉ ketone. Structure activity relationships (SARs) predict that the conjugated double bond system accounts for almost half of the OH reactivity. The conjugated double bond would therefore be expected to have a partial rate coefficient of the order of 1 × 10⁻¹⁰ cm³ molec⁻¹ s⁻¹ (i.e. similar to OH + isoprene) (Atkinson and Arey, 2003b). Consistent with this, the reported yields of acetone and formaldehyde from OH + myrcene are similar (Atkinson and Arey, 2003b), suggesting that the isolated double bond and the conjugated double bond system have comparable OH reactivity, as such we would expect C₉ and C₁₀ co-products to be formed in comparable yields. With a significant fraction of reactions with OH leading to the loss of three carbon atoms from the parent structure, the straight chain monoterpenes gas-phase spectra tend to contain fewer features of MW greater than that of the precursor and more mid-MW features. It tends to be these mid-MW features, such as m/z 111 and 93 (e.g. 4-vinyl-4-pentenal – MRY 1.2 in Fig. 9, MH⁺ and MH⁺-H₂O, respectively) and 113 (e.g. 2-methylenepentanediol, MH⁺; Lee et al., 2006) that assist in the classification of the straight chain monoterpenes experiments within the statistical space. Besides these ions, m/z 139 (primary myrcene C₉ aldehyde and/or C₉ ketone product) also assists in separating the myrcene spectra from those of α-pinene.

By comparing both the gas- and particle-phase PCA results for cyclic monoterpenes in Figs. 2 and 7a, it is evident that the dominant loadings represent compounds of similar MW, i.e. m/z 169, 151 and 107 (primary aldehyde product, e.g. pinonaldehyde- PINAL in Fig. 9, parent ion and fragments thereof) and 139 (primary ketone product parent ion) for the gas phase and m/z 187, 185, 183 and 169 for the particle phase. Conversely, for the straight chain monoterpenes experiments the major gas-phase loadings represent compounds of significantly smaller MW than their particle-phase counterparts, i.e. m/z 113, 111 and 93, compared to m/z 325, 322, 321, 227 and 215. Indeed, the straight chain monoterpenes LC-MS/MS spectra contained on average ~10 % more signal >250 Da than the cyclic monoterpenes spectra. Also, the composition of the ions observed in the straight chain monoterpenes LC-MS/MS spectra suggests that the SOA particles contained both oligomers and highly oxidised species with the C₁₀ backbone intact (i.e. O : C = 0.6), similar in structure to (but a little less oxidised than) extremely low volatility organic vapours (ELV-VOC), which have been observed previously in significant yield from α-pinene and limonene (as well as 6-nonenal) ozonolysis chamber experiments in the absence of an OH scavenger, as well as from boreal forests in Finland (Ehn et al., 2014). Further evidence to elucidate the type of SOA formed from the oxidation of straight chain monoterpenes can be obtained from investigation of the grouping of myrcene spectra in the cTOF-AMS PCA (Fig. 8a). In the hour-4 cTOF-AMS PCA loadings bi-plot, we see that the grouping of the myrcene spectra is influenced somewhat by both m/z 41 and 44, indicating the presence of unsaturated hydrocarbon material and LV-OOA in the SOA, potentially a result of oligomerisation or further oxidative heterogeneous chemistry involving reaction at remaining C=C double bond site(s).

β-caryophyllene readily forms particulate matter on oxidation (e.g. Jaoui et al., 2003; Lee et al., 2006; Winterhalter et al., 2009; Alfarra et al., 2012; Chen et al., 2012), with reaction predominantly at one of the two C=C sites, e.g. with OH or O₃, although O₃ attack occurs almost exclusively at the endocyclic double bond (Jenkin et al., 2012), yielding relatively low vapour pressure, unsaturated and oxygenated primary products (Fig. 9), which have significant affinity for the particle-phase (Jenkin et al., 2012). A further oxidation step involving the second C=C site can result in increased oxygen (and/or nitrogen, depending on NOₓ conditions) content, yet with little, if any reduction in the original C number. As with the cyclic monoterpenes PCAs, the CIR-TOF-MS and LC-MS/MS PCA bi-plots demonstrate similarities in terms of classifying β-caryophyllene oxidation and SOA formation with comparable MW species, e.g. primary products β-caryophyllene aldehyde (MW 236, BCAL in Fig. 9) and β-caryophyllene secondary ozonide in the gas-phase (MW 252, BCSOZ in Fig. 9), β-caryophyllonic acid (MW 252, C141CO2H in Fig. 9) in both phases and secondary product β-nocaryophyllinic acid (MW 254, C131CO2H in Fig. 9) in the particle-phase. In the hour-4 cTOF-AMS PCA loadings bi-plot, the myrcene and β-caryophyllene clusters are located adjacent to one another, with β-caryophyllene classification primarily influenced by the m/z 41 peak, which as described above, is indicative of unsaturated hydrocarbon content (Alfarra et al., 2012).

5.2 Mapping within a class

Within the cyclic monoterpenes group there is a small degree of separation between the limonene and α-pinene experiments, with three out of the four α-pinene experiments located to the upper and right region of the monoterpenes cluster. This distribution/separation within the group may be a consequence of precursor-specific reaction pathways; for instance, although structurally similar, α-pinene and limonene react at somewhat different rates with respect to both OH and O₃ (Atkinson and Arey, 2003a). Over a fixed time period, such system reactivity will govern the degree of oxygenated content present within a closed analyte matrix and may facilitate the isolation of specific reaction pathways. Furthermore, the separation of such similar gas-phase precursors within a class cluster may help us to elucidate differences in resultant SOA yield and composition (e.g. limonene tends to have a larger SOA yield than α-pinene; Lee et al., 2006; Fry et al., 2014). It therefore may be possible with the use of larger and more detailed data sets, to employ loading information to determine the importance of certain products to SOA composi-
5.3 Mapping reactivity

In order to explore how the PCA technique can be used to investigate product distributions driven by certain starting conditions, a separate analysis was conducted on the five toluene experiments. In this instance we investigate the product distribution dependency on initial VOC/NO ratios. The VOC/NO ratios employed nominally represent “low”, “medium” and “high” NO conditions, with values of roughly 11 (i.e. low NO, “NO-limited” NO formation conditions – as determined from simulation chamber ozone isopleth plots; see Wagner et al., 2003; two experiments), 4 (i.e. moderate NO; two experiments) and 1 (i.e. high NO, “VOC-limited”; one experiment), respectively. The resultant PCA loadings bi-plot (produced using the methodology described in Sect. 3) is given in Fig. 10.

From inspection of the PCA loadings bi-plot in Fig. 10, it is clear that the toluene photooxidation spectra distribute in statistical space according to their respective initial VOC/NO ratios. Figure 10 shows the low NO, high VOC/NO ratio experiments grouped in the lower right-hand quadrant of the PCA space, principally influenced by loadings representing toluene (m/z 93 and 77, parent and fragment ions, respectively; note m/z 93 off-scale in Fig. 10) and cresol (m/z 109). Summed spectra containing larger quantities of precursor would suggest the presence of a less reactive environment, which is the case here, where low NO levels in the NO-limited regime, result in low [OH] (reduced radical cycling) and low [O3] (less NO to NO2 conversions) (see also Bloss et al., 2005). Similarly, the relatively low contribution from cresol to the low NO3 summed spectra, originates from a larger net cresol concentration across the experiment on account of low system reactivity (i.e. loss via reaction with OH).

The moderate NO, medium VOC/NO experiments group uniquely in the lower left-hand quadrant of the PCA space in Fig. 10, principally on account of loadings representing benzaldehyde (m/z 107) and the ring-opening products: citraconic anhydride (m/z 113); 4-oxo-2-pentenal, maleic anhydride and/or angelicalactone (m/z 99) and methyl glyoxal (m/z 73). The greater abundance of higher-generation, ring-opening products implies a more reactive environment (i.e. increased chemical processing) than that formed under low NO conditions. Larger net benzaldehyde concentrations are likely to originate from greater system reactivity and greater abundance of NO to fuel the forming RO2 + NO reaction.

The high NO, low VOC/NO ratio experiment is sited in the left-hand half of Fig. 10, on account of it possessing higher system reactivity (with respect to the low NO experiments) and the resultant greater proportion of ring-opening product ions (as is the case for the moderate NO3 experiments). However, the low VOC/NO ratio experiment is uniquely displaced into the upper region of the PCA space owing to a large contribution from 2-butenedial and/or 2(5H)-furanone (m/z 85, off scale in Fig. 10) to the summed spectra (the yields of both of which are likely to be important under high NO conditions, owing to reaction through the RO2 + NO channel).

6 Atmospheric relevance and future directions

Having successfully used the mechanistic fingerprints in the chamber data to construct descriptive statistical models of the gas and particle phases, and having applied the methodology to map mesocosm environments, a next logical step would be to use this detailed chemical framework to investigate ambient VOC and SOA composition data, in an attempt to help elucidate and deconvolve the important chemistry controlling the gas- and particle-phase composition of inherently more complex real-world environments.

If ambient biogenic gas/particle composition spectra of unknown origin, uncertain speciated composition and/or a high level of detail and complexity were to be mapped onto the relevant statistical model (i.e. introduced as a separate test set), their resultant vector description in the statistical space could provide information regarding the type of precursors present and the underlying chemical mechanisms at play, as exemplified by the classifying of the mesocosm experiments by the fraction of isoprene, monoterpenes, and sesquiterpenes.
chemistry in the experimental fingerprints. Furthermore, as shown by the mapping of toluene photooxidation experiments into a separate and distinct cluster, the methodology is potentially able to be robust with respect to other chemical compositions expected for a real-world environment that is significantly impacted by both anthropogenic and biogenic emissions (e.g. Houston, USA and the Black Forest – Munich, DE). This capability is important when attempting to understand the complex interactions that exist between urban and rural atmospheres and when attempting to understand VOC and SOA source identification.

One potential problem in moving from simulation chamber data to real-world systems, would be the applicability of using “static” experimental spectra (i.e. time averaged) to build a model to accept “dynamic” data, in which there would be potentially overlapping reaction coordinates and multiple precursor and radical sources.

In order to investigate the impact of a more dynamic system on the composition of the gas-phase matrix and hence on the composition of the spectra employed to build the model, a zero-dimensional chamber box model was constructed for the α-pinene system and operated under three different scenarios:

1. Basic chamber simulation: α-pinene concentration constrained to measurements (initial concentration 124 ppbV); NO and NO₂ initialised according to measurements (31 and 41 ppbV, respectively).
2. Spiked chamber simulation: α-pinene constrained as in (1), but profile duplicated to represent a fresh injection of the precursor (at the midpoint of the experiment) on top of the already evolving matrix; constant 10 ppbV HONO employed as NO and radical source.
3. Constant injection chamber simulation: α-pinene and HONO constrained to constant values of 5 and 10 ppbV, respectively.

It should be noted here that the model runs are not idealised. The aim of these simulations is only to provide systematically more complex chemical systems with which to compare and contrast a simulation representing the measured data set. For work regarding the evaluation of the MCM with respect to single VOC precursor chamber experiments (including model-measurement intercomparison), see, for example, Bloss et al. (2005) (toluene), Metzger et al. (2008) and Rickard et al. (2010) (1,3,5-TMB), Camredon et al. (2010) (α-pinene) and Jenkin et al. (2012) (β-caryophyllene).

The results of the three different model scenarios are given in Fig. 11, mapped through to the resultant simulated mass spectra (i.e. integrated across the experiment).

Figure 11a and b show the results from scenario (1). Figure 11a gives the evolution of the respective systems over the molecular weight region of interest with time; colour scale represents relative abundance (%). Right-hand plots represent (b) relative abundance of simulated molecular weights during straight α-pinene photooxidation; (d) difference in relative abundance of simulated molecular weights between double injection of α-pinene continuous HONO input and straight α-pinene photooxidation; (f) difference in relative abundance of simulated molecular weights between continuous α-pinene and HONO input and straight α-pinene photooxidation. See text for details.

Figure 11c and d show the results from scenario (2). Figure 11c clearly shows the second α-pinene injection on top of the evolving matrix and the resultant system evolution. Figure 11d shows the “difference model mass spectra” between scenarios (1) and (2), from which it can clearly be seen that there is very little difference between the spectra of the basic model and the “spiked” system. The difference in “mass channel” relative abundance (ΔMC) is generally ≤ ± 2%,
with the exceptions of MWs 168 and 186. MW 168 primarily comprises pinonaldehyde, with a ΔMC of around −6%; pinonaldehyde is a primary product and is slightly lower in relative abundance in scenario (2) owing to the longer reaction time employed and the greater proportion of pinonaldehyde reacted. MW 186 comprises a number of primary and secondary products and has a ΔMC of roughly +3%.

The results from model scenario (3) are given in Fig. 11e and f. As with scenario (2), there is no dramatic difference between the simulated mass spectra of scenario (3) and the base-case scenario (1). In this instance ΔMC is generally ≤±5%, with the exceptions of MWs 136 and 168 and MWs 121 and 245. The relative abundance of the precursor is lower in this case on account of the constraining method employed and once again the relative abundance of pinonaldehyde is slightly lower owing to the longer reaction time. MW 121 solely comprises PAN and MW 245 primarily comprises a C$_{10}$ tertiary nitrate (C$_{10}$H$_{15}$NO$_6$, MCM designation: C106NO3). Both species are slightly elevated with respect to the base-case scenario (3) owing to the longer reaction time and the continual input of OH and NO into the model in the form HONO.

Scenarios (2) and (3) represent complex mixtures with overlapping reaction coordinates, each one step closer to a real-world case than scenario (1) and the chamber data employed within this work. However, despite the increase in complexity of the scenarios, both exhibit very little compositional difference to the base-case scenario and hence the chamber data employed in this work. These results give some confidence that despite being constructed from summed simulation chamber data, the statistical models employed here represents a solid framework onto which real atmosphere spectra could be mapped and interpreted.

A further step in increasing complexity and hence a further step towards the real-world system, would be the addition of other (potentially unidentified) precursors to the simulation, which may be at different stages of oxidation or have passed through different reactive environments. Further increases in complexity, beyond the analysis discussed here, will form the focus of future work.

7 Conclusions

A chemometric dimension reduction methodology, comprising PCA, HCA and PLS-DA has been successfully applied for the first time to complex gas- and particle-phase composition spectra of a wide range of BVOC and mesocosm environmental simulation chamber photooxidation experiments. The results show that the oxidised gas-phase atmosphere (i.e. the integrated reaction coordinate) of each different structural type of BVOC can be classified into a distinct group according to the controlling chemistry and the products formed. Indeed, a potential major strength of the data analysis methodology described here, could lie in the decoding of mechanisms into pathways (i.e. separation within a group on account of different underlying chemistry) and consequently linking chemical pathways to precursor compounds. Furthermore, the methodology was similarly able to differentiate between the types of SOA particles formed by each different class of terpene, both in the detailed and broad chemical composition spectra. In concert, these results show the different SOA formation chemistry, starting in the gas-phase, proceeding to govern the differences between the various terpene particle compositions.

The ability of the methodology employed here to efficiently and effectively “data mine” large and complex data sets becomes particularly pertinent when considering that modern instrumentation/techniques produce large quantities of high-resolution temporal and speciated data over potentially long observation periods. Such statistical mapping of organic reactivity offers the ability to simplify complex chemical data sets and provide rapid and meaningful insight into detailed reaction systems comprising hundreds of reactive species. Moreover, the demonstrated methodology has the potential to assist in the evaluation of (chamber and real-world) modelling results, providing easy to use, comprehensive observational metrics with which to test and evaluate model mechanisms and outputs, and thus help advance our understanding of complex organic oxidation chemistry and SOA formation.

The Supplement related to this article is available online at doi:10.5194/acp-15-8077-2015-supplement.

Acknowledgements. The authors gratefully acknowledge the UK Natural Environment Research Council (NERC) for funding the APPRAISE ACES consortium (NE/E011217/1) and the TRAPOZ project (NE/E016081/1); the EU-FP7 EUROCHAMP-2 programme for funding the TOXIC project (E2-2009-06-24-0001); the EU ACCENT Access to Infrastructures program for funding work at the PSI and the EU PEGASOS project (FP7-ENV-2010-265148) for funding used to support this work. A. R. Rickard and M. R. Alfarra were supported by the NERC National Centre for Atmospheric Sciences (NCAS). The authors would like to thank the University of Leicester Atmospheric Chemistry group for assistance throughout all experiments, including Alex Parker, Chris Whyte, Iain White and Timo Carr; co-workers at the University of Manchester for assistance with MAC experiments; co-workers from Fundacion CEAM, Marie Camredon and Salim Alam for assistance with EUPHORE experiments and co-workers from the Laboratory of Atmospheric Chemistry smog chamber facility at the Paul Scherrer Institute (PSI) for assistance with PSISC experiments. The authors are grateful to M. Wiseman from the University of Brighton, for useful discussions regarding various statistical techniques.

Edited by: V. F. McNeill
References


